A FEATURE FUSION METHOD BASED ON EXTREME LEARNING MACHINE FOR SPEECH EMOTION RECOGNITION

Lili Guo[1], *Longbiao Wang*[1,*], *Jianwu Dang*[1,2,*], *Linjuan Zhang*[1], *Haotian Guan*[1,3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Intelligent Spoken Language Technology (Tianjin) Co. Ltd., Tianjin, China

ABSTRACT

Speech emotion recognition is important to understand users' intention in human-computer interaction. However, it is a challenging task partly because we cannot clearly know which feature and model are effective to distinguish emotions. Previous studies utilize convolutional neural network (CNN) directly on spectrograms to extract features, and bidirectional long short term memory (BLSTM) is the state-of-the-art model. However, there are two problems of CNN-BLSTM. Firstly, it doesn't utilize heuristic features based on priori knowledge. Secondly, BLSTM has a complex structure and high complexity in training. To address the first problem, we propose a feature fusion method that combines CNN-based features and heuristic-based discriminative features which are extracted from heuristic features using deep neural network (DNN). In addition, we utilize extreme learning machine (ELM) instead of BLSTM to solve the second problem. The experiments conducted on EmoDB and our method leads to 40% relative error reduction in F1-score compared to CNN-BLSTM.

Index Terms— speech emotion recognition, convolutional neural network (CNN), extreme learning machine (ELM), bottleneck features, heuristic features

1. INTRODUCTION

Human-computer interaction has become prevalent in spoken dialogue systems and intelligent voice assistants, etc. Speech emotion recognition can significantly help machines to understand users' intention, so accurately distinguish users' emotion can provide great interactivity. However, it is a difficult task because we cannot clearly know which feature and model are effective to distinguish emotions [1]. In addition, there is no unified way to express emotions, so features should have good robustness for different express ways.

Conventional strategies for speech emotion recognition are selecting heuristic features (such as MFCC, pitch, energy, etc.) based on human priori knowledge [2]. The most commonly used model is first to obtain fixed-length segment-level features and then to train a method such as BLSTM to get the utterance-level label [3]. [4] proposed DNN-ELM model which utilized deep neural network (DNN) to obtain emotion state probability distribution. And then, a simple classifier extreme learning machine (ELM) was used to get labels. [5] made improvements about the DNN-ELM model. It used the activations of the last hidden layer of DNN instead of probability distributions to train ELM. [6] proposed recently the recurrent neural network (RNN)-ELM model which accounted for long contextual effect in emotional speech. However, it is difficult to select effective features just based on priori knowledge, and some priori knowledge is not very accurate. In addition, it will take much time in selecting features.

To address above problems, convolutional neural network (CNN) was used to extract features [7]. In recent years, CNN has been applied in speech area and shows great performance [8, 9]. [10] utilized CNN to extract features from spectrograms, and then SVM was trained as a classifier. [11] and [12] proposed a hybrid CNN-BLSTM model directly on spectrograms, and CNN-BLSTM has become the state-of-the-art approach at present. Yet many problems still exist in this approach. 1) In aspects of features, it does not utilize heuristic features based on priori knowledge, so it cannot effectively use heuristic-based features which are useful for speech emotion recognition. 2) In aspects of models, the framework of BLSTM is complicated, and it needs lots of training data. When train data is insufficient, it is easy fall into overfitting.

To address the first problem, we propose a feature fusion method that combines CNN-based features and heuristic-based discriminative features. It is the first work to combine them in speech emotion recognition task. As raw heuristic features are correlate and can only reflect critical value which leads to small inter-class distance [13], we extract bottleneck features from the raw heuristic features by using DNN. Bottleneck features force the information pertinent to classification into a low-dimensional representation which is discriminative. We then combine bottleneck features and CNN-based features. To address the second problem, we use ELM instead of conventional BLSTM to distinguish emotions. ELM has been applied in many classification tasks because of prop-

*Corresponding author: longbiao_wang@tju.edu.cn; jdang@jaist.ac.jp.

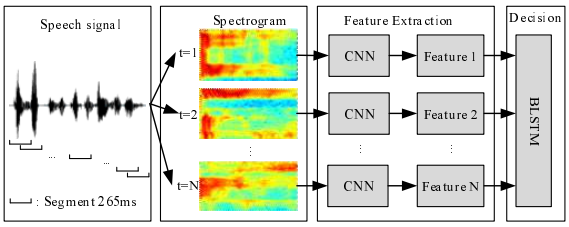


Fig. 1. Structure of CNN-BLSTM.

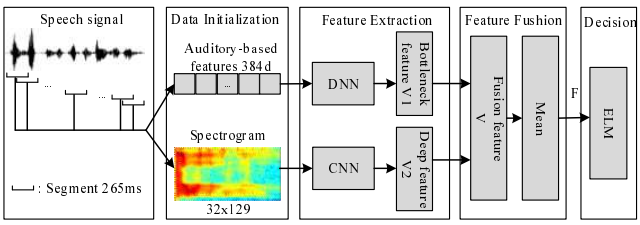


Fig. 2. Structure of our speech emotion recognition method.

erties of high generalization capability and fast training. In addition, ELM performs well on small dataset. As far as we know, CNN-ELM is firstly proposed and applied in speech emotion recognition task.

The rest of this paper is organized as follows. Section 2 introduces the baseline model CNN-BLSTM. We then describe our method in Section 3. Experiments are conducted in Section 4. Section 5 makes conclusions and prospects.

2. BASELINE MODEL

BLSTM is widely used in time-distributed fields [14]. The main idea of BLSTM is utilizing forward direction LSTM and backward LSTM to extract the hidden information in future and past, and the two parts of information forms the final output. It can utilize the context information which is important in speech field. Fig. 1 shows the structure of CNN-BLSTM. Firstly, speech signal is divided into N segments with fixed length. Then it transform speech signal to spectrogram by short time Fourier Transform (STFT). When STFT, we use the default values of 256 FFT points, 256 window size and 50% overlap. CNN is used to extract segment-level features from spectrogram. Finally, those features feed to BLSTM to get utterance-level label.

There are two main problems of this model. Firstly, it does not utilize heuristic features based on priori knowledge. Secondly, BLSTM has a complicated structure, and there are many parameters need to been adjusted.

3. FEATURE FUSION METHOD BASED ON ELM

Our method makes improvements both in aspects of features and in aspects of models, which is shown in Fig. 2. In aspects of features, we add heuristic-based features to CNN-based features. In this work, we use feature fusion rather than decision fusion. Feature fusion is simpler than decision fusion, and is easy to design, so it was adopted in many systems [15]. In addition, feature fusion only needs one decision method without considering the weight between two decision methods. We do not fuse those features directly but extract the discriminative bottleneck features $V1$ from the raw heuristic-based features using DNN. CNN-based features $V2$ extracted from segment-level spectrograms using CNN. Before feature fusion, normalization is needed because different maximum and minimum values will weaken fusion effects. Bottleneck features and CNN-based features are all normalized to range of 0-1. Then we combine the bottleneck features with the CNN-based features in one large feature vectors V and the feature vectors of t-th segment in i-th utterance as following formula:

$$V_i^t = \left[V1_i^t, V2_i^t\right] \tag{1}$$

where $V1_i^t$ is the heuristic-based discriminative feature vectors of t-th segment in i-th utterance, and $V2_i^t$ is the CNN-based feature vectors of t-th segment in i-th utterance.

We then use ELM instead of BLSTM to distinguish emotions. Because all the utterances have been divided into fixed-length segment, the segment-level features cannot feed to ELM directly. We perform mean operation to all the segments in one utterance to get feature F_i of i-th utterance.

$$F_i = \frac{1}{n_i}\sum_{t=1}^{n_i} V_i^t \tag{2}$$

where n_i is segment number of i-th utterance.

3.1. Heuristic-based Discriminative Feature Extraction

This subsection will introduce how to extract bottleneck features $V1$. As heuristic-based features are correlate and can only reflect critical value that will lead to small inter-class distance, so it is necessity to extract discriminative feature.

Firstly, we use 265 ms window size and 25 ms window shift to transform an utterance into segments. Although the window size is a challenge problem, researchers have found that a segment speech signal which is greater than 250 ms includes competent emotional information [4]. Then, we use openSMILE [16] tool to obtain the heuristic features, and then compute segment-level statistical features with 384 dimensions which proposed in [17].

Bottleneck features are extracted by DNN in which one hidden layer has less hidden units than other hidden layers. The hidden layer with less hidden units transforms the emotional information into a low-dimensional representation. Deep belief network (DBN) has been proved to be able to use its deep structure to model actual signals in nature [18], so we use DBN to model DNN. The training of DBN adopts the unsupervised greed algorithm which uses the stack RBM for

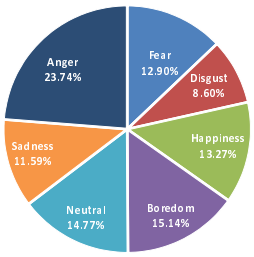


Fig. 3. Emotion distribution of EmoDB.

pre-training [19]. After pre-training, all parameters of DBN in each layer are taken as the initial parameters of DNN, and then the error back propagation (BP) algorithm is used to fine tuning. Heuristic-based features with 384 dimensions used as inputs. After many attempts, we choose the hidden units of DNN with 100, 30 and 100. The outputs of the second hidden layer are the bottleneck features $V1$ with 30 dimensions.

3.2. ELM-based Decision

As BLSTM has a complex structure and slow training speed. In addition, BLSTM may not be well trained when data is insufficient. To address this problem, we choose ELM as the classifier instead of BLSTM. ELM proposed by Huang is a learning algorithm for single-hidden layer feed-forward neural networks (SLFNs) [20]. Its advantages are high generalization capability and fast training [21]. We only need to set the number of hidden layer units and doesn't need to adjust parameters. Its training process finished in one time without any iterations, which leads to faster than conventional BP-based algorithms such as BLSTM. Furthermore, ELM performs well on small database, so the fusion features V are fed to ELM for emotion classification in utterance level.

4. EXPERIMENTS

4.1. Experimental Setup

We take experiments on the famous EmoDB [22]. It consists of 535 utterances in German with seven emotions. All utterances are sampled at 16 KHz with approximate 2-3 seconds long. Fig. 3 shows the emotion distribution of EmoDB. We can see that anger holds the highest percentage of 23.74, while disgust holds the lowest percentage of 8.60. This imbalance is a typical problem for many databases. As it is a small database we adopt 10-fold cross-validation in experiments.

We experimented with different numbers of hidden units and layers, learning rate, etc. Finally, we choose the optimal structure among all attempts. When training the CNN and DNN, all segments in one utterance share the label, and we choose cross entropy as the cost function. The structure of CNN contains two convolutional layers and two max-pooling layers. The first convolutional layer has 32 filters with 5×5 size, and the second convolutional layer has 64 filters with 5×5 size. The pooling size of two pooling layers is 2×2. We adopt a full connected layer with 1024 units. In order to avoid over-fitting, a dropout layer with 0.5 factor is used before output layer. All the experiments list as follow:

DNN-ELM [5]: Its inputs are segment-level heuristic-based features with 384 dimensions. There are four hidden layers in DNN, each with 512 units.

CNN-BLSTM [12]: As shown in Fig. 1, it uses CNN directly on spectrograms to extract acoustic features. It uses two hidden layers BLSTM, each with 200 units.

CNN-BLSTM (+heuristic features): It makes improvement in aspects of features. We make a combination of CNN-based features and heuristic-based discriminative features. Then the segment-level fusion features feed to BLSTM. We use two hidden layers BLSTM, each with 200 units.

CNN-ELM: This method makes improvement in aspects of models. The number of hidden units in ELM is 2100.

CNN-ELM (+heuristic features): This is our method as shown in Fig. 2. Hidden layer units of ELM is 2100.

Table 1. F1 (%) comparison of bottleneck features and raw heuristic features.

Emotion class	Raw heuristic Features	Bottleneck Features	Change
Fear	67.74	66.67	-1.07
Disgust	79.07	80.43	+1.36
Happiness	60.94	68.66	+7.72
Boredom	73.94	76.02	+2.08
Neutral	69.82	83.87	+14.05
Sadness	84.03	82.26	-1.77
Anger	80.29	85.28	+4.99
Average	**73.69**	**77.60**	**+3.91**

4.2. Validation of Bottleneck features

We take experiments to validate the discriminative bottleneck features. Because ELM is adopted as the classifier in our method, we also use ELM method in this part. As ELM is a static classifier, we first perform mean operation to all the segments in each utterance to get utterance-level features. Then the raw heuristic features with 384 dimensions and the bottleneck features with 30 dimensions are fed to ELM independently. Table 1 illustrates the F1 score which is a most common used measure of a test's accuracy.

Results in Table 1 show that there is a great improvement when using bottleneck features. Bottleneck features can enhance the outperformance 3.91% on average F1 compared with raw heuristic features, especially for neutral class (+14.05%). The results prove that it is necessity to extract discriminative bottleneck features.

Table 2. Comparison of emotion recognition models.

Model	Precision(%)	Recall(%)	F1(%)
DNN-ELM	85.55	84.09	84.56
CNN-BLSTM	84.91	86.66	87.49
CNN-BLSTM (+heuristic features)	90.22	87.73	89.68
CNN-ELM	92.64	90.83	91.47
CNN-ELM (+heuristic features)	93.30	91.97	92.50

4.3. Evaluation Results

Table 2 lists the mean results of seven emotions in terms of precision, recall and F1. From the table we can draw the conclusions: 1) CNN-BLSTM (+heuristic features) outperforms CNN-BLSTM by over 18% relative error reduction in terms of F1, proving that heuristic features based on priori knowledge can enhance emotion recognition. 2) CNN-ELM significantly outperforms CNN-BLSTM with about 31% relative error reduction in terms of F1, indicating that ELM as decision method is more effective than BLSTM in this task. We assume that it is partly because CNN has extracted effective features, so those utterances can be easily classified by a static classifier. In addition, ELM can perform well on small dataset. 3) Results of CNN-ELM (+heuristic features) are better than other models. One reason is the combination of heuristic-based discriminative features and CNN-based features. Another reason is that our model adopts ELM as decision maker. It outperforms DNN-ELM by 51% relative error reduction and outperforms CNN-BLSTM by around 40% relative error reduction in F1. By comparing CNN-ELM (+heuristic features) and CNN-ELM we can see that fusion feature leads to over 12% relative error reduction in F1.

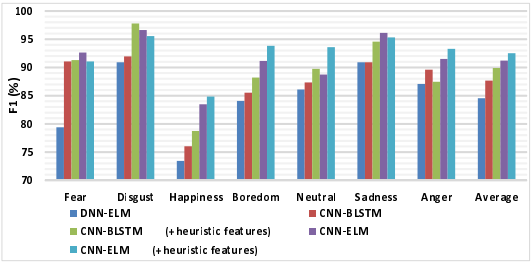
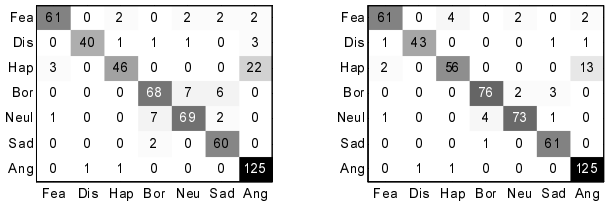


Fig. 4. F1 results for each emotion.

Fig.4 shows the F1-score for each emotion. It is found: 1) CNN-ELM (+heuristic features) achieves best performance in classes of happiness, boredom, neutral and anger. However, when inferring utterances labeled fear, disgust and sadness, its results are not the best. The reason might be that utterances labeled fear, disgust and sadness hold a low proportion. 2) For fear and sadness utterances CNN-ELM performs best, which indicates that the proposed hybrid model is effective.



(a) CNN-BLSTM (b) Our method

Fig. 5. Confusion matrices of CNN-BLSTM and our method.

CNN-BLSTM (+heuristic features) obtains best result in class of disgust, indicating the effective of fusion features. 3) In average F1, our methods CNN-BLSTM (+heuristic features), CNN-ELM and CNN-ELM (+heuristic features) get better results than other models. Overall, the proposed method is effective in emotion recognition task.

To analyze the relation between each emotion, Fig. 5 shows the confusion matrices of the proposed method and CNN-BLSTM. The abscissa as detected labels, and the ordinate as actual labels. 1) One can see that many confusions is concentrated between happiness and anger. In Fig.5(a) about 30% happiness utterances are detected as anger. Although our method makes great improvement, there are still has some confusions in Fig. 5(b). We assume that it is because both happiness and anger have high values of energy and arousal. However, there are under 1% anger utterances detected as happiness. We assume the reason is that as anger utterances hold the highest percentage. 2) In addition, there are many confusions between boredom and neutral. Fig. 5(a) shows about 8.6% boredom utterances detected as neutral and 8.8% neutral utterances detected as boredom. It may be because both boredom and neutral are peaceful mood and low arousal value. Our method weakens this confusion in Fig. 5(b).

5. CONCLUSION

We proposed a feature fusion method that combines CNN-based features and heuristic-based discriminative features. And then, ELM was fed to distinguish emotions. As far as we know, our work is the first one to combine heuristic features with CNN-based features for speech emotion recognition. In addition, it is the first work that combines CNN and ELM in speech emotion recognition task. Experiment results indicate that the proposed method has encouraging performance. Although automatic feature gets great results, heuristic features still have assignable contribution.

6. ACKNOWLEDGEMENTS

The research was supported by the National Natural Science Foundation of China (No. 61771333 and No. U1736219), JSPS KAKENHI Grant (16K00297) and Didi Chuxing.

7. REFERENCES

[1] M.E. Ayadi, M.S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] F. Eyben, M. Wllme, and B. Schuller, "Openearintroducing the munich open-source emotion and affect recognition toolkit," in *3rd international conference on Affective computing and intelligent interaction and workshops*, 2009, pp. 1–6.

[3] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5688–5691.

[4] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of INTERSPEECH*, 2014, pp. 223–227.

[5] Z.Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5150–5154.

[6] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition.," in *Proceedings of INTERSPEECH*, 2015, pp. 1537–1540.

[7] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5115–5119.

[8] Ebrahimi K.S., V. Michalski, K. Konda, et al., "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 467–474.

[9] T.N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 4580–4584.

[10] Z.W. Huang, M. Dong, Q.R. Mao, and Y.Z. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[11] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Signal and Information Processing Association Annual Summit and Conference*, 2016, pp. 1–4.

[12] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proceedings of INTERSPEECH*, 2017, pp. 1089–1093.

[13] F. Grezl and P. Fousek, "Optimizing bottle-neck features for lvcsr," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4729–4732.

[14] G. Keren and B. Schuller, "Convolutional rnn: an enhanced model for extracting features from sequential data," in *International Joint Conference on Neural Networks*, 2016, pp. 3412–3419.

[15] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2539–2544.

[16] F. Eyben, M. Wllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[17] B. Schuller, S. Stefan, and B. Anton, "The interspeech 2009 emotion challenge," in *Proceedings of INTERSPEECH*, 2009.

[18] D. Yu and M.L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proceedings of INTERSPEECH*, 2011, pp. 237–240.

[19] G. Hinton, L. Deng, D. Yu, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[20] G.B. Huang, Q.Y. Zhu, and C.K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[21] Q.Y. Zhu, A.K. Qin, P.N. Suganthan, and G.B. Huang, "Evolutionary extreme learning machine," *Pattern recognition*, vol. 38, no. 10, pp. 1759–1763, 2005.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, et al., "A database of german emotional speech," in *Proceedings of INTERSPEECH*, 2005, vol. 5, pp. 1517–1520.